Journal of Soft Computing Exploration

Analysis of public opinion sentiment against COVID-19 in Indonesia on twitter using the k-nearest neighbor algorithm and decision tree

Ryo Pambudi[1], Faiq Madani[2]
[1,2]Department of Information System, Universitas Diponegoro, Indonesia

Article Info

Article history:

Received Sep 21, 2022
Revised Sep 24, 2022
Accepted Sep 29, 2022

Keywords:

Sentiment analysis
Classification
Public opinion
Public sentiment

ABSTRACT

COVID-19 has become an ongoing disease pandemic across the globe. The need for information makes social media such as twitter a place to exchange information. This tweet can be used to see public sentiment towards COVID-19 in Indonesia. Sentiment analysis classifies opinions from tweets that have been processed and classified into different sentiments, namely negative, neutral, or positive. The aim of this paper is to find the algorithm that has the best accuracy. The researcher proposes to compare the K-Nearest Neighbors (KNN) and decision tree algorithms to be used in the classification of sentiment data from tweets related to COVID-19 that took place in Indonesia. The results of the evaluation of performance metrics concluded that the decision tree algorithm has a higher level of accuracy than KNN. Decision tree produces accuracy = 0.765, error = 0.235, recall = 0.76, and precision = 0.767 which is better when compared to KNN which produces accuracy = 0.69, error = 0.31, recall = 0.66, and precision = 0.702.

Corresponding Author:

Ryo Pambudi,
Department of Information System,
Universitas Diponegoro,
Imam Bardjo SH No. 5, Semarang, Indonesia.
Email: ryopambudi@students.undip.ac.id

1. INTRODUCTION

The COVID-19 pandemic in Indonesia is part of the 2019 coronavirus disease (COVID-19) outbreak caused by severe acute respiratory coronavirus syndrome 2 (SARS-CoV-2) defined by the World Health Organization (WHO) as a pandemic global event that has taken place around the world [1], [2]. The COVID-19 pandemic is putting pressure on the entire world, and it has already claimed thousands of lives in the different pandemic-affected nations [3]. Its impact has spread in various sectors of life, not only in the health sector but also in the economy, government, higher education [4] and nternational travels were suspended [5].

Artificial intelligence contributes to effective text processing and extracting information regarding human concerns with respect to patients and cases of death during and after the pandemic that can determine epidemic prevention COVID-19 [6]. Social media has a significant impact on daily life because it links users to the outside world [7]. Social media can be used as a source of public opinion on various aspects of controversial issues such as COVID-19 [8]. COVID-19 is becoming one of the important topics or trends that are of concern to the public on various social media. Twitter is one of the social media that contains sentiments that can be obtained easily and quickly [9]. With more than 200 million active users and 10.6 billion tweets sent worldwide. Twitter itself has grown to be one of the most widely used social media platforms [10]. Twitter limits the number of characters in tweets and posts. This makes tweets or posts posted by users contain messages that are shorter, denser, and clearer so that users are more concise in providing

information or opinions [11]. The freedom to take advantage of the Twitter social network offers users the opportunity to write tweets, comments, or feedback expressing their information or opinions about the COVID-19 pandemic situation [12]. This tweet can be used to see public sentiment towards the handling of COVID-19.

Sentiment analysis related to public opinion is useful for aligning opinions for the community in assessing an issue or relevant information. Sentiment analysis aims to find opinions about certain entities that are based on and assessed at the level of documents, sentences, or words [13], [14]. Sentiment analysis from Twitter should focus on classification problems, where inputs on sentiment mining are classified as positive or negative [15]. Positive sentiment expresses a good opinion in a context; negative sentiment expresses a bad opinion in a context; while neutral sentiment states things that do not support good or bad. The text mining technique is used in the process of sentiment analysis [16]. Text mining takes place by analyzing large amounts of text from tweets, finding patterns, and extracting possible information from the text [17]. The algorithms used in conducting sentiment analysis in this study are KNN and decision tree to find out which classifier gives the best results in terms of accuracy, precision, and recall. The results of this study will provide an overview to the general public on whether COVID-19 tends to a positive or negative opinion as well as compare the accuracy of the two KNN algorithms and Decision Tree.

2. METHOD

In this paper, the research methodology is carried out through several stages, including identification of problems and solutions, proposed methods, and finally evaluation and conclusions. The steps taken are used to process Twitter data related to COVID-19 in Indonesia to get the results of sentiment analysis classified using the decision tree algorithm and KNN. Figure 1 shows the research methodology in this study.

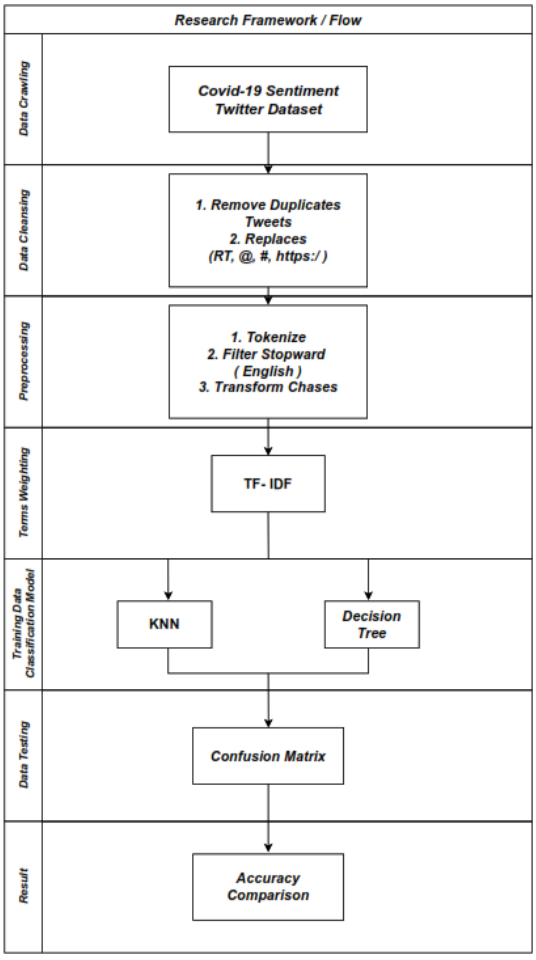

Figure 1. Research method flow

2.1. Problem identification and solution

At the stage of identifying problems and solutions, researchers identify problems with Twitter users who provide comments or comments. At this stage, a literature study is carried out regarding the selection of algorithms and how to apply them. Next, determine the solution to the problem. Finally, the aim of the study is to compare the results with the accuracy of the two K-Nearest-Neighbors (KNN) and decision tree algorithms regarding public opinion sentiment regarding COVID-19 in Indonesia.

2.2. Pre-processing data

In the proposed method stage, the researcher collects data from Twitter. After classification, then preprocessing is carried out on the data , including data cleansing, folding cases, tokenization, stop words removal, and stemming Fitri and colleagues (2019).

- Data cleansing is the cleaning of irrelevant tweet data so that it becomes relevant data.
- Folding case is the process of changing words into the same form, for example, lowercase or uppercase letters.
- Tokenization means dividing the sentence into several parts called tokens. Tokens can be formed from words, phrases, or other meaningful elements.
- Stop word removal is the removal of words that are common and often used but do not have a significant effect in sentences. The Twitter messages correspond to a list of stopwords containing stopwords in Indonesian such as (and), (or), etc.
- Stemming , which is the process of getting basic words by removing affixes and suffixes .

2.3. Processing data

Data representation into numerical form, data having to share into training data and test data, and data grouping to determine the variables to be used are all processes involved in data processing [18]. At the data processing stage, researchers took a total of 1,000 data points to be carried out in the classification process. where the data consists of 500 positive sentiments and 500 negative sentiments. Before performing the classification, it is necessary to determine the portion ratio between the training data and the testing data. In this study, the random state was determined to be 0.2, which means that the portion of the comparison between the training and testing data is 20% of the testing data and 80% of the training data. After that, the classification process is carried out using two algorithms, namely KNN and decision tree.

Decision trees are a method that is quite efficient at making data classifiers. The decision tree has a flowchart-like structure that resembles a tree, where each internal node tests an attribute, each branch represents the test results, and each terminal node represents a class label. While the node at the top of the decision tree is the root node [19]. The KNN algorithm itself has a fair amount of accuracy. The KNN algorithm itself has a fair amount of accuracy [20]. The KNN algorithm classifies by comparing unknown data points with similar training data points, measured by the Euclidean distance. Attribute values are normalized to prevent attributes with larger ranges from exceeding attributes with smaller ranges [21]. The Euclidean distance metric calculation is used with the number of neighbors, or the K value, of 3. After the classification of the 2 classification algorithms is carried out, an evaluation is carried out to determine the performance of the KNN algorithm and decision tree.

2.4. Evaluation or conclusion

At this stage, the testing/evaluation process is carried out to determine the performance of each algorithm. In this research, the testing method is 10-fold cross-validation. Cross-Validation (CV) is a statistical method that can be used to evaluate the performance of a model or algorithm where the data is separated into two subsets, namely learning process data and validation/evaluation data. The model or algorithm is trained by the learning subset and validated by the validation subset [22]. Furthermore, the selection of the type of CV can be based on the size of the dataset. The CV k-fold test is used because it can reduce computation time while maintaining the accuracy of the estimate. After conducting testing performance metrics, a confusion matrix for count score accuracy was generated. The confusion matrix can visualize the performance of an algorithm, especially the model algorithm of the classification that is carried out when making a prediction, so that it can provide knowledge not only about errors made in classification but also about errors made. In classifiers, there is always something called a "false statement. For example, a sentence is declared negative even though the sentence is a positive sentence, or a sentence is declared

neutral even though the sentence is a positive/negative sentence, and a positive sentence is declared negative, which can be seen in Table 1.

Table 1. Classification for performance evaluation

Classification Category	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

In evaluating the performance of the metrics, the accuracy, error rate, recall (sensitivity), precision, F-Measure, False Positive Rate (FPR), False Negative Rate (FNR), and specificity values are calculated.

3. RESULTS AND DISCUSSIONS

The data used in this study is a data set from Twitter social media with comments on the COVID-19 discussion in Indonesia. The keywords used are hashtags that are within the scope of the cases discussed. The data used is relevant data that has been cleaned, namely 1000 comments. The data is manually labeled as positive and negative sentiment. There are several examples of manual labeling of tweets with different sentiments, which can be seen in Table 2.

Table 2. Sentiment labeling on tweets

Tweet	Sentiment
Let's work together to help the government break the chain of covid-19 and break the chain of hatred. So that Indonesia can return to normal	Positive
The first is because many Indonesian people underestimate it, the second the government is less firm and less alert in dealing with this covid-19	Negative
Yes, this may be the effect of the government's unpreparedness from the start to prevent the COVID-19 virus from entering Indonesia. Moreover, the handling is still chaotic in areas. Hopefully everything ends and returns to normal	Negative
What Indonesia needs right now is the moral support of all the people for the front groups and the government, both central and regional. God willing, all government decisions can reduce the spread of COVID-19.	Positive
Indonesians are not stressed because they think about covid-19 but think about the government which is not firm and does not have one voice.	Negative
For all Indonesian people. We should stay at home following the advice of the government. To be able to break the chain of covid-19 if we all stay at home consistently. God willing, this plague will end quickly, and we can all welcome the month of Ramadan.	Positive

From the dataset processed with 1000 data consist of 500 negative sentiments and 500 positive sentiments, the data is then split for validation with a ratio of 80:20. Classification is carried out on each of the KNN and Decision Tree algorithms so that the accuracy values of the two algorithms can be compared. From the classification, the TP, FN, FP, and TN values for each algorithm can be seen in Table 3.

Table 3. Classification results

Algorithm	TN	FP	FN	TP
KNN	72	28	34	66
Decision Tree	77	23	24	76

The results of the performance evaluation of the KNN and decision tree algorithms can be seen in Table 4.

Table 4. Results of evaluation of performance metrics

Performance Metrics	KNN	Decision Tree
Accuracy	0.69	0.765
Error rate	0.31	0.235
Recall	0.66	0.76
Precision	0.702	0.767
F-Measure	0.680	0.763
FPR	0.28	0.23
FNR	0.34	0.24
Specificity	0.72	0.77

According to the metric performance evaluation results, the decision tree algorithm has a higher level of accuracy than KNN. The decision tree achieves accuracy = 0.765, error = 0.235, recall = 0.76, and precision = 0.767, which is superior to KNN's accuracy = 0.69, error = 0.31, recall = 0.66, and precision = 0.702. The results of the comparison of the accuracy of the two algorithms can be seen in Figure 2.

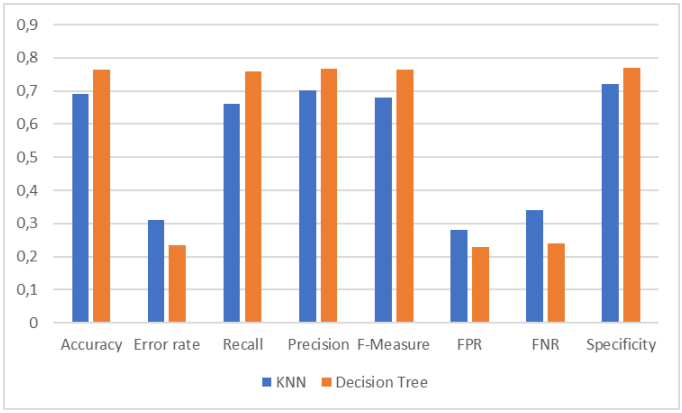

Figure 2. Comparison diagram of accuracy results between KNN and decision tree

4. CONCLUSION

Based on the sentiment dataset from Twitter about COVID-19 in Indonesia by Twitter users from 1000 test data, consisting of 500 comments with positive sentiments and 500 comments with negative sentiments. Here it can be concluded that the decision tree algorithm is considered better and more effective than the KNN algorithm for analyzing public opinion sentiment regarding COVID-19 in Indonesia. Based on the results of the accuracy of the decision tree algorithm, the algorithm has a higher level of accuracy than KNN. The decision tree produces accuracy = 0.765, error = 0.235, recall = 0.76, and precision = 0.767, which is better than KNN, which produces accuracy = 0.69, error = 0.31, recall = 0.66, and precision = 0.702. So it can be concluded that the decision tree algorithm has better accuracy than KNN because it is used for classifying public opinion related to COVID-19 in Indonesia on Twitter social media.

REFERENCES

[1] M. T. Hasan, "The sum of all scares COVID-19 sentiment and asset return," *Q. Rev. Econ. Financ.*, vol. 86, pp. 332–346, 2022.

[2] Z. Lyu and H. Takikawa, "Media framing and expression of anti-China sentiment in COVID-19-related news discourse: An analysis using deep learning methods," *Heliyon*, vol. 8, no. 8, p. e10419, 2022.

[3] R. Djalante *et al.*, "Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020," *Prog. Disaster Sci.*, vol. 6, 2020.

[4] A. V. Pelt, H. A. Glick, W. Yang, D. Rubin, M. Feldman, and S. Kimmel, "Evaluation of COVID-19 testing strategies for repopulating college and university campuses: a decision tree analysis," *J. Adolesc. helath*, vol. 68, pp. 28–34, 2020.

[5] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "COVID-19 vaccine hesitancy : text mining , sentiment analysis and machine learning on COVID-19 vaccination twitter dataset," *Expert Syst. Appl.*, p. 118715, 2022.

[6] Y. Didi, A. Walha, and A. Wali, "COVID-19 tweets classification based on a hybrid word embedding method," *Big Data Cogn. Comput.*, vol. 6, no. 2, p. 58, 2022.

[7] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Appl. Soft Comput. J.*, vol. 97, p. 106754, 2020.

[8] C. Shofiya and S. Abidi, "Sentiment analysis on COVID-19 related social distancing in Canada using twitter data," *Int. J. Environ. Res. Public Health*, vol. 18, no. 11, p. 5993, 2021.

[9] J. X. Koh and T. M. Liew, "How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds," *J. Psychiatr. Res.*, vol. 145, no. October 2020, pp. 317–324, 2022.

[10] D. A. Efrilianda, E. N. Dianti, and O. G. Khoirunnisa, "Analysis of twitter sentiment in COVID-19 era using fuzzy logic method," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 1–5, 2021.

[11] J. Abraham, D. Higdon, and J. Nelson, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," *SMU Data Sci. Rev.*, vol. 1, no. 3, p. 22, 2018.

[12] F. Es-Sabery *et al.*, "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier," *IEEE Access*, vol. 9, pp. 58706–58739, 2021.

[13] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment Analysis of Social Media Twitter with Case of Anti- LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019.

[14] T. Mustaqim, K. Umam, and M. A. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," in *J. Phys.: Conf. Ser.*, 2020, vol. 1567, no. 3, p. 32024.

[15] A. Skuza Michałand Romanowski, "Sentiment analysis of Twitter data within big data distributed environment for stock prediction," *Proc. 2015 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2015*, vol. 5, pp. 1349–1354, 2015.

[16] A. Falasari and M. A. Muslim, "Optimize Naïve Bayes Classifier Using Chi Square and Term Frequency Inverse Document Frequency For Amazon Review Sentiment Analysis," *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 31–36, 2022.

[17] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The Comparation of Text Mining With Naive Bayes Classifier , Nearest Neighbor , and Decision Tree to Detect Indonesian Swear Words on Twitter," *2017 5th Int. Conf. Cyber IT Serv. Manag. (CITSM). IEEE*, pp. 1–5, 2017.

[18] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," *Sci. J. Informatics*, vol. 5, no. 1, p. 18, 2018.

[19] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," *2018 Int. Conf. Orange Technol. ICOT 2018*, no. October, 2018.

[20] A. Susanto, D. Sinaga, C. A. Sari, E. H. Rachmawanto, and D. R. I. M. Setiadi, "A High Performace of Local Binary Pattern on Classify Javanese Character Classification," *Sci. J. Informatics*, vol. 5, no. 1, p. 8, 2018.

[21] S. Hota and S. Pathak, "KNN classifier based approach for multi-class sentiment analysis of twitter data," *Int. J. Eng. Technol.*, vol. 7, no. 3, pp. 1372–1375, 2018.

[22] B. Prasetiyo, M. A. Muslim, and N. Baroroh, "Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique," in *J. Phys.: Conf. Ser.*, 2021, vol. 1918, no. 4, p. 42002.